Search of Sequence Databases with Uninterpreted High-Energy Collision-Induced Dissociation Spectra of Peptides

John R. Yates, III and Jimmy K. Eng
Department of Molecular Biotechnology, University of Washington, Seattle, Washington, USA

Karl R. Clauser and Alma L. Burlingame
Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, USA

We have broadened the utility of the SEQUEST computer algorithm to permit correlation of uninterpreted high-energy collision-induced dissociation data of peptides with all sequences in a database. SEQUEST now allows for the additional fragment ion types observed under high-energy conditions. We analyzed spectra from peptides isolated following trypsin digestion of 13 proteins. SEQUEST ranked the correct sequence first for 90% (18/20) of the spectra in searches of the OWL database, *without* constraint by enzyme cleavage specificity or species of origin. All false-positives were flagged by the scoring system. SEQUEST searches databases for sequences that correspond to the precursor ion mass ±0.5 u. Preliminary ranking of the top 500 candidates is done by calculation of fragment ion masses for each sequence, and comparison to the measured ion masses on the basis of ion series continuity, summed ion intensity, and immonium ion presence. Final ranking is done by construction of model spectra for the 500 candidates and constructing/performing of a cross-correlation analysis with the actual spectrum. Given the need to relate mounting genome sequence information with corresponding suites of proteins that comprise the cellular molecular machinery, tandem mass spectrometry appears destined to play the leading role in accelerating protein identification on the large scale required. © *1996 American Society for Mass Spectrometry (J Am Soc Mass Spectrom 1996, 7, 1089–1098)*

The rapid progress in genome sequencing of a variety of organisms is producing a vast sequence infrastructure intended to facilitate the identification of genes and elucidation of their biological function [1]. To eventually understand gene function–dysfunction, correlation with expression of the corresponding protein through biochemical experimentation at the cell or tissue level is required. There are estimated to be ~ 250 specific cell types in humans, each of which expresses and regulates a subset of the ~ 100,000 human genes as proteins [2]. Hence, understanding such vast numbers of proteins, their functions, and interconversions throughout the life of a cell provides a challenge on a scale much larger than the genome sequencing efforts. Because proteins actually carry out most cellular processes, it is of primary importance to focus on establishing their identity and physiologically active forms. This goal represents a major experimental challenge that is not addressed readily by classical methods of protein sequencing [3]. Although clearly a daunting task, it has recently become apparent that this problem may be addressed by using tandem mass spectrometry to rapidly determine both partial and complete sequence [4–7], despite early concerns about the extent of expertise required to interpret the tandem mass spectra of peptides. Furthermore, sequence information provides far more discriminating power to search databases than peptide-mass fingerprinting approaches [8–12]. Sequence data also substantially increases the success rate and certainty in protein identification experiments to the point where only a single peptide may be necessary (assuming the presence of only a single protein in the sample) [4, 5].

Recent instrumental innovations have brought about the availability of a variety of tandem mass spectrometric methods for the generation of peptide fragmentation information that enables determination, to vary-

Address reprint requests to John R. Yates, Department of Molecular Biotechnology, Box 357730, University of Washington, Seattle, WA 98195-7730 or Alma L. Burlingame, Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94143.

1044-0305/96/$15.00
PII S1044-0305(96)00079-7

Received January 30, 1996
Revised June 21, 1996
Accepted June 24, 1996

ing extents, of a peptide's sequence and accompanying covalent modifications. In general, these methods employ either low-energy (10–100 eV) [13, 14] or high-energy (1000–8000 eV) [15–18] collisions with neutral atoms to induce dissociation of the protonated molecular ion, or analyze the metastable (postsource decay) fragment ions formed from excess energy deposition during matrix-assisted laser desorption ionization (MALDI) [19]. The fragmentation patterns produced by using high-energy collisions, such as those observed on tandem double-focusing [15], double-focusing–orthogonal acceleration time-of-flight (TOF) [16], and collision cell equipped reflectron TOF [18] mass spectrometers contain a wider range of fragment ion types than is observed under low-energy collision-induced dissociation (CID) conditions or MALDI metastable conditions. By using low-energy conditions, the main fragment ions observed are associated with cleavage of the peptide backbone. Such cleavages produce sequence ions that exhibit charge retention either on the N-terminal portion of the peptide (*a*, *b*, and *c* ions) or on the C-terminal portion (*x*, *y*, and *z* ions) [15, 20]. High-energy CID of peptides usually generates additional types of fragments that arise from cleavage of the β- or γ-carbon atoms for certain residues referred to as "satellite" sequence ions [15, 20]. These side-chain cleavages are labeled as either *d* or *v* and *w* ions depending on whether the charge (basic residue) is preferentially located on the N- or C-terminal moiety, respectively. In low-energy CID and MALDI metastable [21] mass spectra, neither *c*, *x*, and *z*-type ions nor side-chain cleavages usually are observed. These additional fragmentation pathways seen by using high-energy collision conditions result in more spectral complexity, but provide more detailed structural information facilitating unambiguous sequence determination. Consequently, computer algorithms have been developed to assist in the analysis of high-energy CID spectra [20, 22–24]. These range from computer algorithms that calculate the expected mass-to-charge ratio values for a given sequence to programs designed to aid in the *de novo* determination of unknown sequences. As long as the peptide is derived from a known protein, database search strategies tend to be more successful than *de novo* interpretation strategies. Inability to discriminate among sequence permutations because of spectral ambiguity hinders *de novo* sequence interpretation, whereas database searches need to discriminate only between the subset of permutations actually present in a database.

In all tandem mass spectrometric methods capable of producing peptide fragmentation information, the interpretation of the spectra obtained is currently the rate-limiting step in eliciting of the desired sequence and structural information. If one seeks to employ automated liquid chromatography–tandem mass spectrometry instrumentation capable of acquiring dozens of tandem mass spectra per hour in the current environment of rapidly growing genomic sequence databases, it is essential to develop strategies that enable rapid and direct interrogation of all known sequences by using "raw" tandem mass spectral data [25–27]. Although Mann and Wilm [28] have developed a powerful method to search protein databases through the use of peptide sequence tags determined from partial manual interpretation of tandem mass spectra yielding as few as two or three sequential amino acids, Eng et al. [25] developed methods that permit the search of protein and nucleotide sequence databases directly by using the *uninterpreted* fragmentation information contained in a whole tandem mass spectrum. The utility of both approaches has been established initially with the low-energy CID spectra (E_{lab} = 10–50 eV) produced by using triple quadrupole-type mass spectrometers that operate under electrospray ionization conditions, where the predominant fragment ions are of the *b* and *y* types. The method of Eng et al. employs a pre-search filter that calculates the user-specified fragment ion types expected for each candidate amino acid sequence in the database within the mass tolerance specified in the search. In this preliminary analysis the ion types from each candidate sequence retrieved from the database are compared to the fragment ions present in the tandem mass spectrum. A preliminary score is calculated based on the summed experimental ion intensity, sequence ion continuity, and the presence of immonium ions and the corresponding amino acid in a candidate sequence. The 500 highest preliminary scoring candidates then undergo further analysis that utilizes a cross-correlation function. In this process a model tandem mass spectrum is constructed for each of these 500 candidate sequences and then compared to the experimental tandem mass spectrum by application of a cross-correlation function to measure the closeness-of-fit [25]. The correct sequence is determined by using a normalized ranking of the cross-correlation scores. Any possible false-positive search results may be discerned (are "flagged") from observation of a small difference (< 0.1) in the normalized cross-correlation score between the first and second ranking sequences.

In the present study, we describe modifications to the Eng et al. [25] computer algorithm SEQUEST that allow the search of protein databases with high-energy CID spectra that may contain fragment ions of the *a*, *b*, *c*, *d*, *v*, *w*, *x*, *y*, and *z* types. The results are presented for 20 high-energy CID spectra. Spectra were chosen from our vast library of high-energy CID spectra obtained on a four-sector instrument to encompass the wide variety of fragmentation patterns and ion types seen in high-energy CID and thus thoroughly evaluate the performance of modifications to SEQUEST. Hence, peptides with a range of masses and that contain basic residues in various positions (including a neutral peptide) were selected. Furthermore, the fragmentation patterns and ion types examined are qualitatively similar to those observed with MALDI on double-focus-

ing–orthogonal acceleration TOF [16] and collision cell equipped reflectron TOF [18] instruments, although these MALDI fragment ions tend to display more abundant neutral-loss ion types.

Experimental

Search Algorithm and Under-Specified Parameters

The original SEQUEST algorithm [25] was modified so that calculation of the following fragment ion types is now enabled: *a*, *b*, *c*, *d*, *v*, *w*, *x*, *y*, *z*, and neutral losses of NH_3 and H_2O from *a*, *b*, and *y* ions [29]. Either all of the ion types or a particular subset may be used. Calculations of the corresponding mass-to-charge ratio values were performed as previously described [15, 29]. The searching method uses the same two-part scoring procedure [25]. Both the preliminary and final scores are calculated with the relationships described in Eng et al. [25], except that the expanded set of ions is included in the calculation. The mass measurement tolerances can be set independently at any value for precursor and fragment ion masses by using either the monoisotopic or average mass scales [29].

These studies employ monoisotopic mass calculations and a high-energy fragment ion set that consists of *a*, *b*, *c*, *d*, *v*, *x*, *y*, *w*, *z*, and neutral loss of NH_3 and H_2O from *b* ions. These ion types are assigned abundances of 1.0, 1.0, 0.5, 1.0, 0.5, 0.5, 1.0, 1.0, 0.5, 0.25, and 0.25, respectively, to approximate their typical relative abundances in high-energy CID spectra and reduce the possibility of false-positives that arise from sequences that correlate implausibly with an experimental spectrum, that is, correlation based on a preponderance of ion types that are generally less common and also less abundant (*c*, *v*, *x*, and *z*) without accompanying major ion types (*a*, *b*, *d*, *w*, and *y*). The same mass tolerances were employed for all searches in this study, namely, ± 0.5 u for the precursor ion and ± 0.0 u for fragment ions. The ± 0.0-u tolerance for fragment ions acts as an effective tolerance of ± 0.5 u because of the way the data are handled for rapid computer calculations. Fragment ion data are evaluated in 1-u bins that are offset by 1.0005 units. The offset ensures that the distribution of possible monoisotopic fragment ion masses is centered in a bin, whereas the ± 0.0-u tolerance assures that no more than one bin is examined, which thus yields a ± 0.5-u tolerance from the center of any given bin. Because $> 95\%$ of the possible peptide fragment ion masses < 1500 u are constrained by elemental composition to windows of width less than ± 0.2 u [30] and fragment ion masses 150–1500 u are measured to ± 0.3 u, a measured mass could be evaluated in an incorrect bin only if its elemental composition resulted in an expected mass that was more than ± 0.2 u from the mean. With the 20 common amino acids such cases are extremely rare; that is, < 0.2 u from the mean requires more than 4 cysteines or more than 6 aspartates, whereas > 0.2 u from the mean requires more than 7 isoleucines, leucines, or lysines.

SEQUEST normally searches for immonium ions as part of the spectral preprocessing routine, but the criteria that requires the presence or absence of particular amino acids in scored sequences based on immonium ion peak detection are conservative. Currently, only H, F, Y, W, and M immonium ions are considered solely because of their usual presence in all types of tandem mass spectra. Because high-energy CID tends to produce near-complete compositional information from a combination of immonium and related ions in the low mass-to-charge ratio range [31] and side-chain loss ions in the high mass-to-charge ratio range [15], SEQUEST was modified to allow amino acid composition or partial peptide sequence information to be user-specified as an additional search constraint. Such constraints may be manually added to the data, which causes the immonium ion portion of the algorithm to be overridden, and thus requires the presence of the specified information in all scored sequences.

The specificity of proteases used to generate a given peptide from its original protein may be included in the search parameters as an option; that is, to narrow the search to consider only those sequences that match the specificity of the protease. In such cases incomplete digestion is allowed to accommodate sequence extensions on either side of a given expected cleavage site.

Databases

The OWL database version 26.0 (105,990 entries) was obtained as an ASCII text file in the FASTA format from the National Center for Biotechnology Information (Washington, DC). OWL is a nonredundant database comprised of protein sequences from the GenBank (Release 88.0, National Center for Biotechnology Information, Washington, DC), SWISS-PROT (Release 31, University of Geneva, Geneva Switzerland), Protein Information Resource (Release 44, National Biology Resource Foundation, Georgetown University, Washington, DC), and National Research Laboratory (Release 18.0, Brookhaven National Laboratory, Brookhaven, NY) databases. Species-specific databases may be constructed by using a computer program to perform keyword searches through a database to transfer all sequences of a particular species to a separate file. Human-specific (17,903 entries) and rat-specific (2505 entries) sequence databases were prepared by extraction of a species-limited set of entries from the complete OWL database. The *S. cerevisiae* sequence database (5252 entries) was obtained from the Stanford yeast sequencing project.

Peptides

The peptides used in these studies (listed in Table 1) were selected from fractions collected from reversed-phase high-performance liquid chromatography sepa-

Table 1. Results of the search of the OWL protein database (release 26; 105,990 sequences) by using high-energy CID spectra

			General search (low-energy ions[a])			General search (high-energy ions[a])				Tryptic search[b] (high-energy ions[a])			
			All species			All species		Single species		All species		Single species	
MH^+	Species[c]	Correct sequence[d]	Top ranking sequence	Rank	ΔC_n[e]	Rank	ΔC_n[e]	Rank	ΔC_n[e]	Rank	ΔC_n[e]	Rank	Δ_n[e]
689.3	H	(K) VWGSIK (G)	KTGWVV	2	0.022	15(1)[f]	0.056	4(1)[f]	0.010	1	0.005	1	0.043
809.4	H	(R) LASYLDK (V)		1	0.007	1	0.009	1	0.041	1	0.171	1	0.306
818.5	H	(K) CamLTAIVK (C)	ACamIATVAV	3	0.062	3(1)[g]	0.015	1	0.135	2(1)[g]	0.006	1	0.383
830.4	H	(K) ILTDKLK (E)	KDTILLAG	> 500	0.000	1	0.013	1	0.051	1	0.088	1	0.171
842.4	R	(K) LPAELATK (Y)	IIALLATQ	3	0.000	1	0.162	1	0.283	1	0.441	1	0.635
904.5	R	(R) VITDLSSGI (–)		1	0.189	1	0.141	1	0.211	1	0.323	1	0.587
996.5	R	(R) VSASDGFWK (S)		1	0.071	1	0.234	1	0.305	1	0.361	1	0.528
1003.5	Y	(K) VTTNINWR	SGGDVGGGGDL	10	0.046	1	0.027	1	0.061	1	0.140	1	0.281
1033.6	R	(R) EFVPPFGIK (G)		1	0.155	1	0.214	1	0.322	1	0.365	1	0.463
1065.5	H	(R) GDFCamIQVGR (N)	TGTAVVDGAFK	10	0.036	1	0.112	1	0.234	1	0.318	1	0.383
1067.5	Y	(K) ESTLHLVLR (L)	ESTLHLVXR	1	0.000	1	0.015	1	0.444	1	0.015	1	0.605
1069.5	H	(Y) RPVAVALDTK (G)	KPVARKASGR	> 500	0.009	1	0.195	1	0.195	—	—	—	—
1081.5	Y	(R) TLSDYNIQK (E)	ATLSDYNLQK	1	0.000	1	0.012	1	0.299	1	0.012	1	0.464
1149.5	H	(K) DRPFFAGLVK (Y)	SALSEFATLNP	61	0.063	1	0.118	1	0.130	1	0.294	1	0.367
1211.6	H	(R) QITVNDLPVGR (S)	KLTITNDHNR	2	0.009	1	0.261	1	0.261	1	0.256	1	0.482
1243.6	H	(R) IQLVEELDR (A)	GEGGAVSRTLTVP	> 500	0.039	1	0.227	1	0.377	1	0.379	1	0.436
1261.9	Y	(R) SDREYPLLIR (M)		1	0.100	1	0.348	1	0.371	1	0.456	1	0.482
1334.6	H	(K) ATAVVDGAFKEVK (L)		1	0.338	1	0.369	1	0.385	1	0.384	1	0.523
1527.8	H	(R) HLREYQDLLNVK (M)	SSTKPSPNNANRVR	> 500	0.094	1	0.369	1	0.438	1	0.419	1	0.569
1533.7	H	(R) KVESLQEEIAFLK (K)		1	0.114	1	0.172	1	0.266	1	0.321	1	0.389

[a]The low-energy ion set included (*b*, *y*, and neutral loss of NH_3, H_2O, and CO from *b* ions). The high-energy ion set included (*a*, *b*, *c*, *d*, *v*, *w*, *x*, *y*, *z*, and neutral loss of NH_3 and H_2O from *b* ions).
[b]Tryptic searches required all sequences to conform to trypsin protease specificity, cleavage after R or K (not RP or KP), but missed cleavages allowed.
[c]H: human, R: rat, Y: yeast.
[d]Parentheses indicate residues before and after peptide. Cam = acrylamide-modified Cys. All searches treated every Cys in the database as acrylamide modified.
[e]ΔC_n: difference in normalized cross-correlation value between the first and second ranked results. A value < 0.1 indicates the first ranked result may be a false-positive.
[f]Rank based on the inclusion of amino acid composition information W derived from immonium and related ions.
[g]Rank based on the inclusion of amino acid composition information Cam derived from neutral losses of the modified side chain from the precursor ion.

rations of tryptic digests of 13 different proteins. The human proteins Cu-Zn superoxide dismutase, cytokeratin (isotype unknown), annexin I (lipocortin I), cytoskeletal tropomyosin, nucleoside diphosphate kinase A, pyruvate kinase, natural killer cell enhancing factor B, fibroblast nonmuscle tropomyosin, and vimentin are from melanoma lysates isolated by two-dimensional gel electrophoresis [5, 32]. The sialyltransferase was purified from rat liver [33]. The yeast signal recognition particle subunits (68 and 19 ku) were purified by immunoaffinity chromatography and electrophoresis [34]. Yeast ubiquitin was obtained from Sigma Chemical Co. (St. Louis, MO).

High-Energy Tandem Mass Spectrometry

All positive-ion high-energy CID spectra were acquired as previously described by using liquid secondary ionization in a thioglycerol matrix on a Kratos Analytical instruments (Columbia, MD) Concept IIHH four-sector mass spectrometer equipped with either a static inlet probe and a 4% electro-optical multichannel array detector [32, 35] or a continuous flow, liquid inlet probe, and a rapid scanning charge-coupled device array detector [5, 36, 37]. Spectra for all of the human peptides except CamLTAIVK were obtained with the instrument in the latter configuration by using 5–50 pmol of material. Spectra for the remaining peptides were obtained with the instrument in the former configuration by using 50–200 pmol of material. Signal-to-noise ratios in all spectra are comparable to the spectra shown in Figures 1 and 2.

Computation

All computer algorithms were written in the C-programming language under the UNIX operating system. All searches were performed on either a Digital Equipment Corporation (Maynard, MA) DEC 3000/700 Alpha Workstation or a Sun Microsystems (Mountain View, CA) SPARC station IPX. A single unconstrained search is completed in less than 2 or less than 12 min, respectively. Search times decrease when species and enzyme cleavage specificity constraints are added.

Results and Discussion

The objective of this investigation was to extend the utility of the database searching algorithm SEQUEST of Eng et al. [25] to include the capability to analyze high-energy CID (E_{lab} = 1–8 keV) spectra of peptides. Following suitable modification, the performance of SEQUEST has been evaluated by performing of searches using 20 high-energy CID spectra of a variety of amino acid compositions and peptide sequences. In these investigations the correct sequence was ranked first in 90% (18 of 20) of unconstrained searches of the entire OWL database as shown in Table 1; 65% (13 of

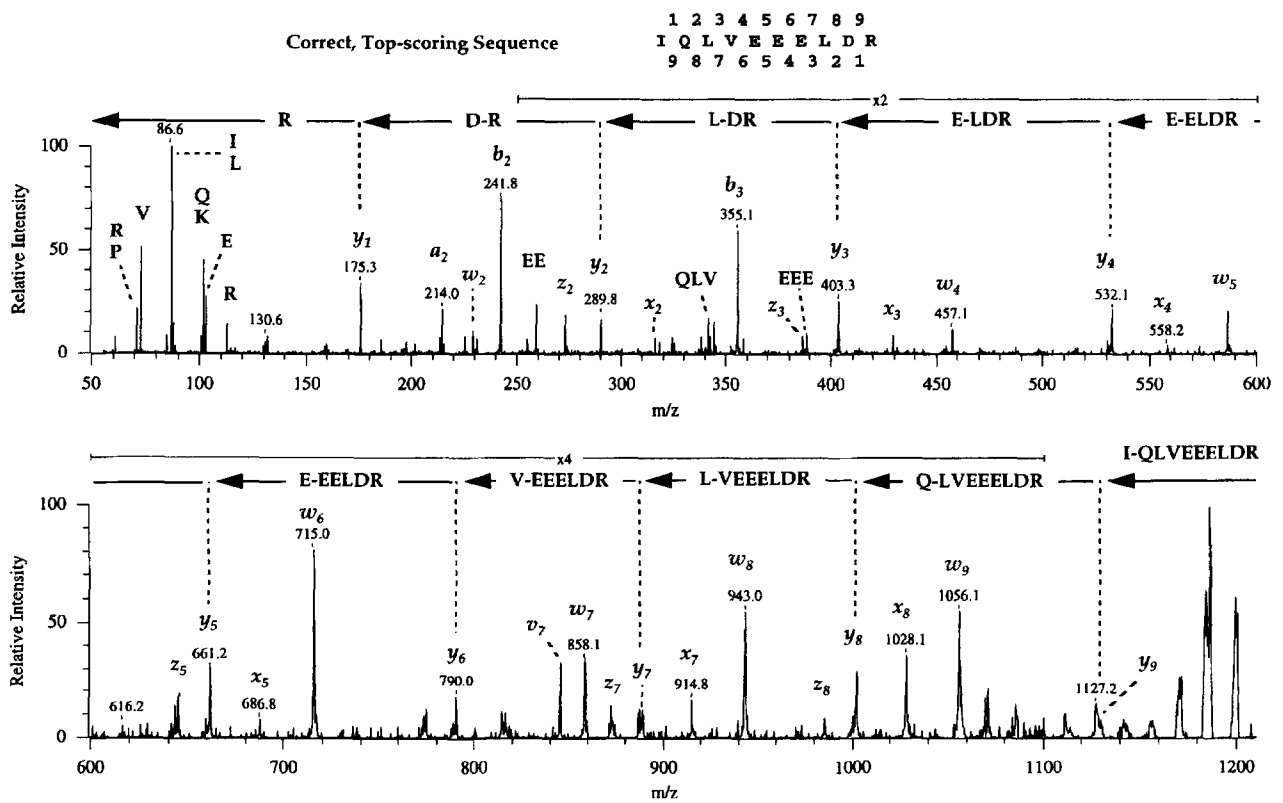

Figure 1. High-energy collision-induced dissociation spectrum obtained from 5–50 pmol of a peptide derived from trypsin digestion of fibroblast nonmuscle tropomyosin isolated from human melanoma cells [4]. MH^+ = 1243.7.

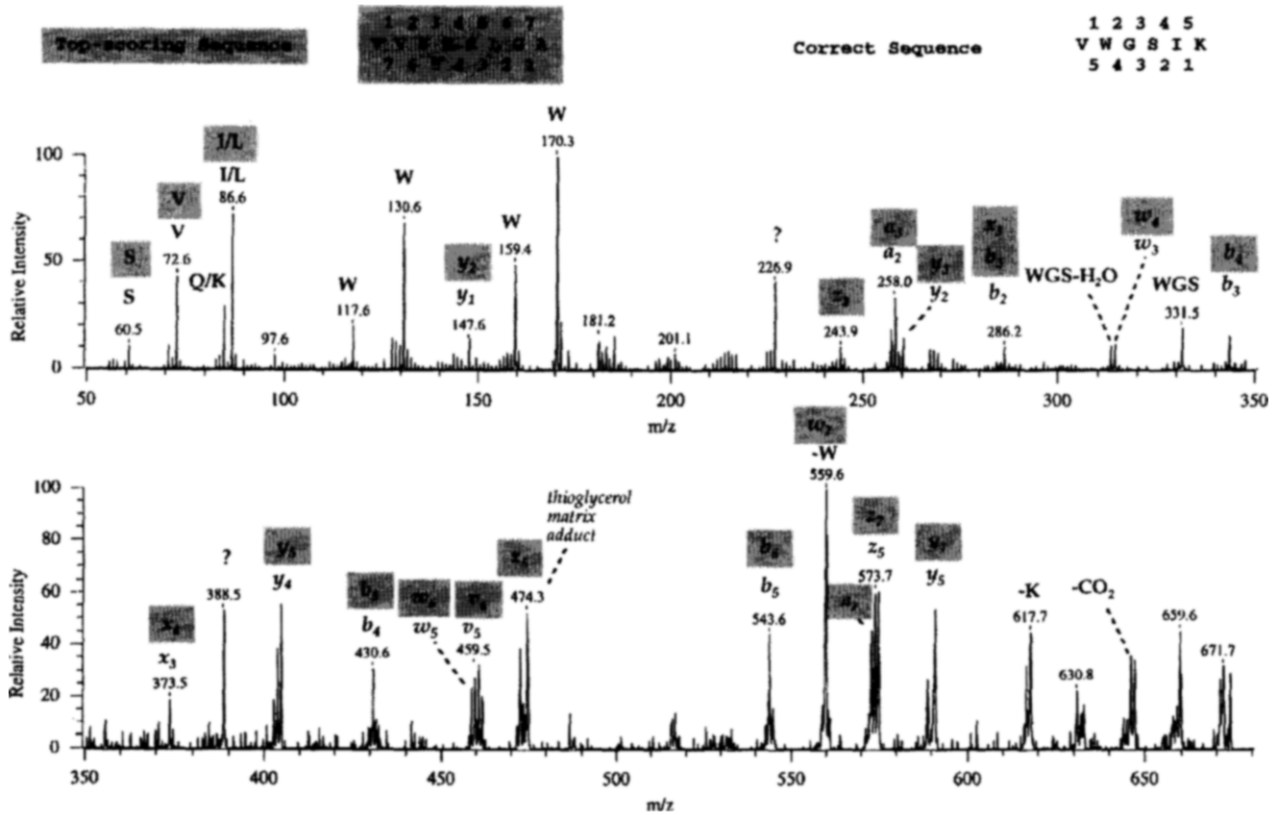

Figure 2. High-energy collision-induced dissociation spectrum obtained from 5–50 pmol of a peptide derived from trypsin digestion of Cu-Zn superoxide dismutase isolated from human melanoma cells [4]. Fragment ion assignments for the top-scoring, but incorrect, sequence are in shaded gray. $MH^+ = 689.3$.

20) of the search results have ΔC_n values > 0.1, which indicates that false-positives are unlikely. Once the searches are limited by enzyme specificity and species of origin the correct sequence is ranked first in 100% (19 of 19) of the searches and 95% (18 of 19) have ΔC_n values > 0.1. The remaining peptide resulted from a nonspecific enzyme cleavage.

Because the high-energy fragmentation behavior of certain types of peptides is qualitatively similar to the low-energy case with just the addition of certain ion types (c, d, v, w, x, z), the low-energy form of the algorithm could be expected to have difficulty scoring the correct sequence from a high-energy CID spectrum. False-positives might arise from incorrect sequences that preferentially fit the data produced by the additional ion types. Following modification of the algorithm to consider the additional ion types that might be present in the data, false-positives should be diminished greatly because the correct sequence could be expected to fit the data much better than any incorrect sequences (unless the data are ambiguous). These expectations are borne out by searches that use the high-energy CID spectrum from the fibroblast nonmuscle tropomyosin peptide IQLVEEELDR (m/z 1243.6), shown in Figure 1, which provides a textbook example of the kind of fragmentation patterns frequently observed under high-energy CID conditions. When this particular spectrum was employed in an unconstrained search of the OWL database (105,990 nonredundant entries) by using SEQUEST in a *low-energy* form (that is, an ion set consisting of b, y, and neutral loss of NH_3, H_2O, and CO from b ions) with the same mass tolerances as the *high-energy* form, the correct sequence was not even ranked among the top 500 sequences in the preliminary search. As shown in Table 1, when the low-energy form of SEQUEST is applied to all 20 of the test spectra for unconstrained searches, only 45% (9 of 20) give the correct result ranked first and only 25% (5 of 20) yield an indication that false-positives are unlikely with ΔC_n values > 0.1. Consequently, the high-energy modifications to the algorithm produced the correct result with substantially stronger scores and substantially reduced the possibility that results could be false-positives.

It is important to note that solely on the basis of the information present in the CID spectrum shown in Figure 1, the N-terminal order of IQ cannot be differentiated from LQ, IK, LK, QL, or KL. This particular ambiguity results from the inability of side-chain loss ions to be formed from an N-terminal amino acid and the isobaric pairing of I/L and Q/K. The additional isobaric permutations QI and KI can be eliminated as possibilities by the presence of the w_9 ion, which allows us to distinguish between I and L at the second

position but not between Q and K. Also, possibilities that contain K are unlikely due to the paucity of ion series with N-terminal charge retention (*a*, *b*, *c*, and *d*). However, in this example none of these theoretical possibilities complicates the search because only the IQ-containing sequence exists in the database.

Spectra That Are Not Correlated Readily

Inspection of Table 1 reveals that despite high-energy modifications to SEQUEST, 2 out of 20 (10%) unconstrained searches of the entire OWL database still did not rank the correct sequence first. However, in both of these searches the small ΔC_n values raise the real possibility of these being false-positive rankings.

One of these spectra is shown in Figure 2. It is the high-energy CID spectrum obtained from the Cu-Zn superoxide dismutase peptide VWGSIK (m/z 689.3). In this example the correct sequence was ranked fifteenth in an unconstrained search with $\Delta C_n = 0.056$. To illustrate the factors involved to obtain such a result, the fragment ion assignments for both the correct sequence and the top scoring sequence VVSGSLGA are shown in Figure 2. This example illustrates the difficulty of using a CID spectrum where the spectrum may not uniquely correlate to a sequence in the database, particularly if some sequence ions are either weak or missing. Searches with spectra from short peptides more frequently encounter this problem, because the database is more likely to contain all the combinatorially possible sequences of amino acids of a given length. However, this spectrum contains a near-complete set of amino acid composition ions. In particular, ions at m/z 117, 130, 159, and 170 together are demonstratively characteristic of the presence of tryptophan in a peptide. When this Trp composition information is included in the search to override the algorithm's conservative immonium ion detection routine and require all scored sequences to contain tryptophan, the correct sequence is ranked first, whereas three other sequences—VVSWAQ, VWSVAGA, and VWSGKL—could be considered as false-positives because their ΔC_n scores are within 0.1 of the correct sequence. Because the peptide used to obtain this spectrum was derived from trypsin digestion of a protein isolated from human melanoma cells, this additional information can be used to limit the sequences considered in a search. Table 1 shows that the correct sequence ranks fourth, first, and first when only sequences are considered that are human, trypsin cleavages of proteins from all species, or trypsin cleavages of human proteins, respectively. In each case, the likelihood of a false-positive result is flagged by ΔC_n scores < 0.1.

Evaluation of the results from the other unconstrained search in which the algorithm ranks an incorrect sequence first, reveals high rankings for sequences with significant structural similarity. An unconstrained search of the entire OWL database with the CID spectrum for the annexin I (lipocortin I) peptide CamLTAIVK (m/z 818.5, where Cam is acrylamide-modified Cys with a modified mass, MW = 174) [32] ranks the incorrect sequences ASETALVK and ADTTAIVK first and second, respectively. The sequence ADTTAIVK also ranks first in a trypsin-specific search, whereas the correct sequence ranks second. The negligible difference in ΔC_n score of these three sequences is a result of the near sequence identity in the last five residues in all three sequences. However, the only ions present in the spectrum that allow unambiguous assignment of CamLTVAIK over the other two are ions that represent unusual losses of −118.0, −105.1, −72.1, and −44.1 from the precursor ion. These ions indicate fragmentation in the side-chain (— C_4H_8NOS, — C_3H_7NOS, — C_3H_6NO, and — CH_2NO, respectively) characteristic of an acrylamide-modified cysteine residue [32]. We typically find acrylamide-modified cysteine residues in proteins, such as this one, which are isolated by sodium dodecylsulfate–polyacrylamide gel electrophoresis [5, 32]. The ability of SEQUEST to consider such side-chain losses from the precursor ion is not currently incorporated. Nonetheless, inclusion of composition information, Cam, in a general search to constrain the sequences considered enables the algorithm to rank CamLTVAIK first (LCamAITVQ and CamLTALVK rank second and third, respectively, with ΔC_n scores < 0.04). Furthermore, if the species of origin of the protein without compositional information is added to the search, then the correct sequence ranks first unambiguously in both general and trypsin-specific searches, because no possible false-positives are indicated by the ΔC_n scores.

Spectra with Potential False-Positive Identifications

Although 18 out of 20 general searches ranked the correct sequence first once SEQUEST was modified to include high-energy ion types, the possibility of a false-positive identification was raised in 5 of these searches by ΔC_n scores < 0.1. In two of these five searches the correct isomer of the isobaric pair Leu and Ile (both have residue mass 113) was identified when sequences were found whose only difference was Leu, Ile, or X (represents an ambiguity between Ile and Leu). In the character representations of amino acids in sequence databases the letter code X signifies any unidentified amino acid. However, in the search algorithm the character X is considered to be either Leu or Ile, but *d* and *w* ions, which can distinguish Leu or Ile, are not scored in favor of X. For example, in a search with the high-energy CID spectrum from the ubiquitin peptide ESTLHLVLR (m/z 1067.5) the correct answer was ranked first, over a second ranking sequence of ESTLHLVXR. The correct answer had both a higher preliminary score and cross-correlation score that identified one more fragment ion, the w_2 ion for leucine. Because the $\Delta C_n = 0.015$, this sequence choice was flagged as a possible false-positive. In the search with

the CID spectrum from the ubiquitin peptide TLSDYNIQK (m/z 1081.5), the correct sequence ranked first and the second ranking sequence was found to be TLSDYNLQK. Because the $\Delta C_n = 0.012$ the result is again flagged as a possible false-positive, but the presence of a w_3 ion for isoleucine allows the correct sequence to be confirmed. Both of these examples demonstrate the exquisite sensitivity possible with this method to search the protein database and the level of structural information present in a high-energy CID mass spectrum. For the purpose of protein identification it is noteworthy that for both of these ubiquitin peptides the isomeric potential false-positives are from the same protein but are derived from different species.

Three other unconstrained searches ranked the correct sequence first, but also indicated the possibility of false-positive identifications with ΔC_n scores < 0.1. The sequence ranking obtained from searches that used spectra that represented the sequences LASYLDK, ILTDKLK, and VTTNINWR (m/z 809.4, 830.4, and 1003.5, respectively) had 5, 13, and 7 other sequences, respectively, with ΔC_n scores within 0.1 of the correct sequence. The five other sequences in the search with the cytokeratin peptide LASYLDK search were from LASYLSR, SPSYLDK, LASYLRS, LASYLKD, and SPSYLEN. These sequences not only have differences from the correct sequence either in the first two amino acids or in the last two amino acids, but also have incorrect pairs of amino acids that are isobaric with the correct pairs. Because a_1, b_1, and c_1 ions are typically not present in high-energy CID spectra, the correct sequence LASYLDK is scored higher based on the presence of only one or two additional fragment ions in the spectrum. The results of the search with the CID spectrum from the cytoskeletal tropomyosin peptide ILTDKLK similarly suffer from the combinatorial possibilities created by amino acids present in the peptide that are isobaric with another amino acid: three residues are either leucine or isoleucine, and two residues are lysine (isobaric with glutamine). The results of the search with the CID spectrum from the 68-ku signal recognition particle subunit peptide VTTNINWR contain seven sequences with ΔC_n scores within 0.1 of the correct sequence due to the inordinately large number of fragment ions in the spectrum. The spectrum not only contains complete or near-complete series of fragment ions from the a, b, v, w, x, y, and z families, but also contains numerous internal acyl fragment ions [15, 29] (the search algorithm currently does not consider internal acyl ions). Taken together, these attributes enable several similar sequences to score well in relation to the correct sequence. However, it is important to note that for all three peptides LASYLDK, ILTDKLK, and VTTNINWR, once information that concerns the species from which the proteins were isolated and the fact that the peptides were derived by tryptic digestion are included in the search parameters, there is no indication of possible false-positive results because the ΔC_n scores are all > 0.1.

Conclusions

These studies have substantiated that the peptide sequence information inherent in high-energy CID spectra is highly specific and can be used to identify amino acid sequences uniquely in the protein database. The resulting power of discrimination is such that 90% (18 of 20) of our spectra were correlated by SEQUEST to the correct sequence in a database of 105,990 nonredundant protein sequences (OWL, release 26) *without* the need to provide the cleavage-specificity of the enzyme used to generate the peptide nor the species from which the original protein was isolated. It is clear that incorporation of additional diagnostic features from high-energy CID spectra that are used routinely in manual interpretation will extend the usefulness to processed and modified peptides such as those that occur as putative MHC antigens and peptides that bear posttranslational or xenobiotic modifications. However, the high success rate observed thus far has some important implications. Search strategies need not be hindered by the occurrence of nonspecific cleavages that result from protease infidelity. Sequence conservation across species allows a tandem mass spectrum to be correlated to a homologous protein from a species other than the one being studied, particularly when the protein is unknown in the species being studied. Among cross-species functional protein homologs sequence identity is typically $> 25\%$, and we commonly find completely conserved tryptic peptides because there tend to be clusters of consecutive sequence identity. Furthermore, the detailed fragmentation information present in a high-energy CID spectrum may allow the presence of a modification to be detected as part of spectral preprocessing to establish criteria for the search algorithm to locate the specific modification sites. Finally, explicit and accurate identification of the location and nature of polymorphic variations in a protein [38], as well as tolerance of sequence errors in the database, will be possible with future modifications to SEQUEST because of the discriminating power of the information in a tandem mass spectrum.

The current and expected future size of sequence databases also raises important issues. A large majority of the combinatorially possible linear sequences of amino acids, greater than eight residues in length, are not currently represented in protein databases. Moreover, the sheer enormity of the combinatorial possibilities (20^n, for peptide length n) combined with the fact that sequences of proteins with common function from evolutionarily related species often have sequence similarity suggests that this will continue to be the case not only when genomes are completely sequenced, but also as genome sequencing is extended beyond those

projects currently underway. Even though a tandem mass spectrum may not be unique for a single amino acid sequence when all the combinatorially possible amino acid sequences are considered, it may very well be unique to a sequence within the genome of a single species. Until the genome from an investigator's species of interest is completely sequenced, however, database searches typically will be conducted through the entire collection of known sequences from all species. Under such conditions, it is likely that subtle differences between conserved sequences may be observed such as the exchange of Leu and Ile seen in the foregoing ubiquitin examples (m/z 1067.5 and 1081.5 in Table 1). In this case the information necessary to differentiate the two isomers is present in the high-energy CID spectra, and SEQUEST ranks the correct sequence first and flags the sequence similarity of the two ubiquitin. Other uncertainties in sequence assignment presented by weak or missing sequence ions in the tandem mass spectrum or isobaric amino acid residues and amino acid combinations in similar sequences may be inconsequential when only one sequence present in a database uniquely fits the spectrum. In the event sequence similarities exist in the complete database that cannot be resolved by the detailed fragmentation information present in the high-energy CID spectra, more restrictions, such as restraining a search to a single species or utilizing proteolytic cleavage information, may provide an unambiguous answer.

The fragmentation patterns produced by the variety of mass spectrometric techniques now available could be analyzed best by allowing consideration of all types of fragment ions. In this manner tandem mass spectra of peptides obtained by employing different internal energy deposition conditions could be used with the algorithm to screen sequence databases readily in an automated fashion with the need for manual evaluation only in the limited number of cases where the algorithm flags the search result as a possible false-positive.

Acknowledgments

This work was supported by the National Science Foundation, Science and Technology Center Cooperative (agreement BIR8809710), National Institutes of Health, National Center for Research Resources (grant RR01614), a University of California Systemwide Biotechnology Grant, and Digital Equipment Corporation. We would like to thank Dr. Katalin F. Medzihradszky and Dr. Steven C. Hall for providing some of the tandem mass spectra, Dr. Diana M. Smith for providing the human melanoma proteins, Fred Walls for technical expertise in acquiring all the tandem mass spectra, and Dr. Ashley McCormack and Max Robinson for insightful discussions.

References

1. Olson, M. *Proc. Nat. Acad. Sci. USA* **1993**, *90*, 4338–4344.
2. Geneser, F. *Textbook of Histology*; Munksgaard: Copenhagen, 1986.
3. Patterson, S. D. *Anal. Biochem.* **1994**, *221*, 1–15.
4. Hunt, D. F.; Henderson, R. A.; Shabanowitz, J.; Sakaguchi, K.; Michel, H.; Sevilir, N.; Cox, A. L.; Apella, E.; Engelhard, V. N. *Science* **1992**, *255*, 1261–1263.
5. Clauser, K. R.; Hall, S. C.; Smith, D. M.; Webb, J. W.; Andrews, L. A.; Tran, H. U.; Epstein, L. B.; Burlingame, A. L. *Nat. Acad. Sci. USA* **1995**, *92*, 5072–5076.
6. Patterson, S. D.; Aebersold, R. *Electrophoresis* **1995**, *16*, 1791–1814.
7. Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. *Nature* **1996**, *379*, 466–469.
8. Yates, J. R.; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993**, *214*, 397–408.
9. Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. *Proc. Nat. Acad. Sci. USA* **1993**, *90*, 5011–5015.
10. James, P.; Quadroni, M.; Carfoli, E.; Gonnet, G. *Biol. Mass Spectrom.* **1993**, *22*, 338–345.
11. Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993**, *22*, 338–345.
12. Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3*, 327–332.
13. Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Nat. Acad. Sci. USA* **1986**, *84*, 620–623.
14. Kaiser, R. E., Jr.; Cooks, R. G.; Syka, J. E. P.; Stafford, G. C. *Rapid Commun. Mass Spectrom.* **1990**, *4*, 30–33.
15. Medzihradszky, K. F.; Burlingame, A. L. *Methods: A Companion to Methods Enzymol.* **1994**, *6*, 284–303.
16. Medzihradszky, K. F.; Adams, G. A.; Burlingame, A. L.; Bateman, R. H.; Green, M. R. *J. Am. Soc. Mass Spectrom.* **1996**, *7*, 1–10.
17. Biemann, K. *Ann. Rev. Biochem.* **1992**, *61*, 977–1010.
18. Carr, S. A.; Roberts, G.; Annan, R. S.; Hemling, M. E.; Hoyes, J. *Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics*; Atlanta, GA, 1995; p 620.
19. Kaufmann, R.; Kirsch, D.; Spengler, B. *Int. J. Mass Spectrom. Ion Processes* **1994**, *131*, 355–385.
20. Johnson, R. S.; Biemann, K. *Biomed. Environ. Mass Spectrom.* **1989**, *18*, 945–957.
21. Rouse, J. C.; Yu, W.; Martin, S. A. *J. Am. Soc. Mass Spectrom.* **1995**, i6, 822–835.
22. Hines, W. M.; Falick, A. M.; Burlingame, A. L.; Gibson, B. W. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 326–336.
23. Papayannopoulos, I. A.; Biemann, K. *J. Am. Soc. Mass Spectrom.* **1991**, *2*, 174–177.
24. Scarberry, R. E.; Zhang, Z.; Knapp, D. R. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 947–961.
25. Eng. J.; McCormack, A. L.; Yates, J. R., III *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
26. Yates, J. R., III; Eng, J.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426–1436.
27. Yates, J. R., III; Eng, J.; McCormack, A. L. *Anal. Chem.* **1995**, *67*, 3202–3210.
28. Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390–4399.
29. Biemann, K. *Methods Enzymol.* **1990**, *193*, 886–888.
30. Mann, M. *Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics*; Atlanta, GA, 1995; p. 639.
31. Falick, A. M.; Hines, W. M.; Medzihradszky, K. F.; Baldwin, M. A.; Gibson, B. W. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 882–893.
32. Hall, S. C.; Smith, D. M.; Masiarz, F. R.; Soo, V. W.; Tran, H. U.; Epstein, L. B.; Burlingame, A. L. *Proc. Nat. Acad. Sci. USA* **1993**, *90*, 1927–1931.

33. Wen, D. X.; Livingston, B. D.; Medzihradszky, K. F.; Kelm, S.; Burlingame, A. L.; Paulson, J. C. *J. Biol. Chem.* **1992**, *267*, 21011–21019.
34. Brown, J. D.; Hann, B. C.; Medzihradszky, K. F.; Niwa, M.; Burlingame, A. L.; Walter, P. *EMBO J.* **1994**, *13*, 4390–4400.
35. Walls, F. C.; Baldwin, M. A.; Falick, A. M.; Gibson, B. W.; Kaur, S.; Maltby, D. A.; Gillece-Castro, B. L.; Medzihradszky, K. F.; Evans, S.; Burlingame, A. L. In *Biological Mass Spectrometry*; Burlingame, A. L.; McCloskey, J. A., Eds.; Elsevier: Amsterdam, 1990; pp 197–216.
36. Walls, F. C.; Hall, S. C.; Medzihradszky, K. F.; Yu, Z.; Burlingame, A. L.; Evans, S.; Hoffman, A. D.; Buchanan, R.; Glover, S. *Proceedings of the 41st ASMS Conference on Mass Spectrometry and Allied Topics*; San Francisco, CA, 1993; pp 937a–937b.
37. Burlingame, A. L. In *Biological Mass Spectrometry Present and Future*; Matsuo, T.; Caprioli, R. M.; Gross, M. L.; Seyama, Y., Eds.; Wiley: Chichester, 1994; pp 147–164.
38. Gibson, B. W.; Biemann, K. *Proc. Nat. Acad. Sci. USA* **1984**, *81*, 1956–1960.